Perspective

When Does Sora Show: The Beginning of TAO to Imaginative Intelligence and Scenarios Engineering

By Fei-Yue Wang, *Fellow, IEEE*, Qinghai Miao, *Senior Member, IEEE*, Lingxi Li, *Senior Member, IEEE* Qinghua Ni, Xuan Li, Juanjuan Li, Lili Fan, Yonglin Tian, and Qing-Long Han, *Fellow, IEEE*

DURING our discussion at workshops for writing "What Does ChatGPT Say: The DAO from Algorithmic Intelligence to Linguistic Intelligence" [1], we had expected the next milestone for Artificial Intelligence (AI) would be in the direction of Imaginative Intelligence (II), i.e., something similar to automatic words-to-videos generation or intelligent digital movies/theater technology that could be used for conducting new "Artificiofactual Experiments" [2] to replace conventional "Counterfactual Experiments" in scientific research and technical development for both natural and social studies [2]–[6]. Now we have OpenAI's Sora, so soon, but this is not the final, actually far away, and it is just the beginning.

As illustrated in [1], [7], there are three levels of intelligence, i.e., Algorithmic Intelligence, Linguistic Intelligence, Imaginative Intelligence, and according to "The Generalized Godel Theorem" [1], they are bounded by the following relationship:

$$\text{AI} \ll \text{LI} \ll \text{II}.$$

Where AlphaGo was the first milestone of Algorithmic Intelligence while ChatGPT that of Linguistic Intelligence. Now with Sora is emerging as the first milestone of Imaginative Intelligence, the triad forms the initial technical version of the decision-making process outlined in Chinese classic *I Ching* (or *Book of Changes*, see Fig. 1): Hexagrams (Rule and Composition), Judgements and Lines (Hexagram Statements and Line Statements, or Question and Answer), and Ten Wings (Commentaries, or Imagination and Illustration).

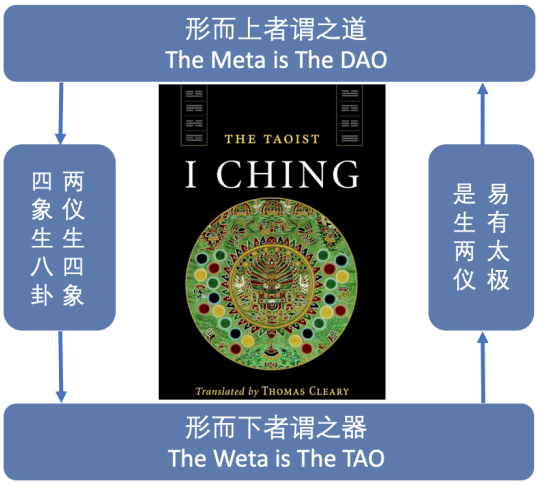

Fig. 1. I Ching: The Book of Changes for Decision Intelligence.

What should we expect for the next milestone in intelligent science and technology? What are their impacts on our life and society? Based on our previous reports in [8], [9] and recent developments in Blockchain and Smart Contracts based DeSci and DAO for decentralized autonomous organizations and operations [10], [11], several workshops [12]–[16] have been organized to address those important issues. The main results have been summarized in this perspective.

Citation: F.-Y. Wang, Q. Miao, L. Li, Q. Ni, X. Li, J. Li, L. Fan, Y. Tian, and Q.-L. Han, "When Does Sora Show: The Beginning of TAO to Imaginative Intelligence and Scenarios Engineering," *IEEE/CAA J. Autom. Sinica*, vol. 11, no. 4, pp. 809-815, Apr. 2024.

F.-Y. Wang is with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China, and with the State Key Laboratory for Management and Control of Complex Systems, Chinese Academy of Sciences, Beijing 100190, China, and also with Faculty of Innovation Engineering, Macau University of Science and Technology, Macau 999078, China (e-mail: feiyue.wang@ia.ac.cn).

Q. Miao, the corresponding author, is with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: miaoqh@ucas.ac.cn).

L. Li is with the Department of Electrical and Computer Engineering, Indiana University-Purdue University Indianapolis, Indianapolis, IN 46202, USA (e-mail: LL7@iupui.edu).

Q. Ni is with the Faculty of Innovation Engineering, Macau University of Science and Technology, Macao 999078, China (e-mail: 3230001810@student.must.edu.mo)

X. Li is with Virtual Reality Fundamental Research Laboratory, Department of Mathematics and Theories, Peng Cheng Laboratory, Shenzhen 518000, China (e-mail: lix05@pcl.ac.cn).

J. Li is with State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: juanjuan.li@ia.ac.cn).

L. Fan is with the School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China (e-mail: lilifan@bit.edu.cn).

Y. Tian is with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China, and also with the State Key Laboratory for Management and Control of Complex Systems, Chinese Academy of Sciences, Beijing 100190, China (e-mail: yonglin.tian@ia.ac.cn).

Q.-L. Han is with School of Science, Computing and Engineering Technologies, Swinburne University of Technology, Melbourne, VIC 3122, Australia (e-mail: qhan@swin.edu.au).

Digital Object Identifier 10.1109/JAS.2024.124383

Historic Perspective

Text-to-Image (T2I) and Text-to-Video (T2V) are two of the most representative applications of Imaginative Intelligence (II). In terms of T2I, traditional methods such as VAE and GAN have been unsatisfactory, prompting OpenAI to explore new avenues with the release of DALL-E in early 2021. DALL-E draws inspiration from the success of language models in the NLP field, treating T2I generation as a sequence-to-sequence translation problem using a discrete variational auto-encoder (VQVAE) and Transformer. By the end of 2021, OpenAI's GLIDE introduced Denoising Diffusion Probabilistic Models (DDPMs) into T2I generation, proposing classifier-free guidance to improve text faithfulness and image quality. The Diffusion Model, with its advantages in high resolution and fidelity, began to dominate the field of image generation. In April 2022, the release of DALL-E2 showcased stunning image generation performance globally, a giant leap made possible by the capabilities of the diffusion model. Subsequently, the T2I field saw a surge, with a series of T2I models developed such as Google's Imagen in May, Parti in June, Midjourney in July, and Stable Diffusion in August, all beginning to commercialize, forming a scalable market.

Compared to T2I, T2V is a more important but more challenging task. On one hand, it is considered important because the model needs to learn the structure and patterns hidden in the video, similar to how humans understand the world through their eyes. Therefore, video generation is a task close to human intelligence and is considered a key path for achieving general artificial intelligence. On the other hand, it is considered difficult because video generation not only needs to learn the appearance and spatial distribution of objects but also needs to learn the dynamic evolution of the world in the temporal domain. In addition, the lack of high-quality video data (especially text-video paired data) and the huge demand for computing power pose great challenges. Therefore, compared to the success of T2I, progress in T2V moves slower. Similar to early T2I, T2V in its initial stages is also based on methods such as GAN and VAE, resulting in low-resolution, short-duration, and minimally dynamic videos that do not reach practical levels.

Nevertheless, the field of video generation has rapidly evolved during the last two years, especially since late 2023, when a large number of new methods emerged. As shown in Fig. 2, these models can be classified according to their underlying backbones. The breakthrough began with language models (Transformer), which fully utilize the attention mechanism and scalability of Transformers; later, the diffusion model family became more prosperous, with high definition and controllability as its advantages. Recently, the strengths of both Transformer and diffusion models have been combined to form the backbone of DiT [17].

The families based on language models are shown on the left side of Fig. 2. VideoGPT [18] utilizes VQVAE for learning discrete latent representations of raw videos, employing 3D convolutions and axial self-attention. A GPT-like architecture models these latents with spatiotemporal position encodings. NUWA [19], an autoregressive encoder-decoder Transformer,

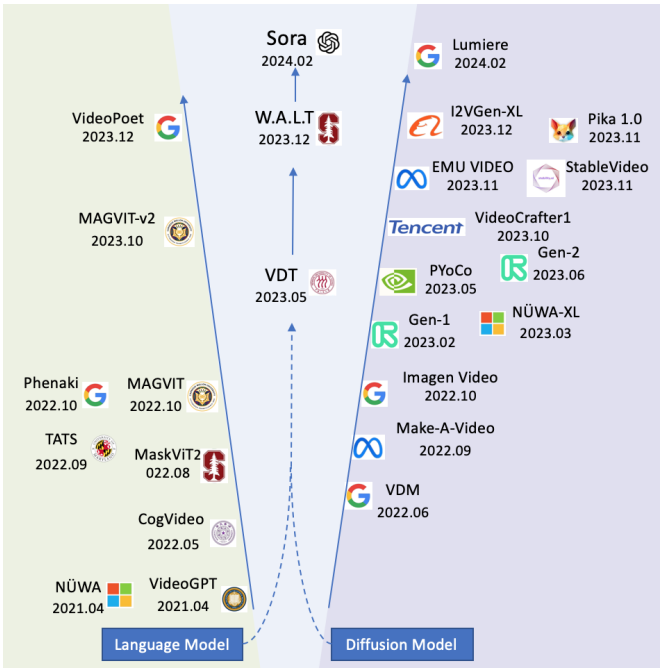

Fig. 2. Brief history of video generation and representative models. Sora indicates the beginning of the Imaginative Intelligence new era.

introduces 3DNA to reduce computational complexity, addressing visual data characteristics. CogVideo [20] is featured as a dual-channel attention Transformer backbone, with a multi-frame rate hierarchical training strategy to better align text and video clips. MaskViT [21] shows that we can create good video prediction models by pre-training transformers via Masked Visual Modeling (MVM). It introduces both spatial and spatiotemporal window attention, as well as a variable percentage of tokens masking ratio. TATS [22] focuses on generating longer videos. Based on 3D-VQGAN and transformers, it introduces a technique that extends the capabilities to produce videos in thousands of frames. Phenaki [23] is a bidirectional masked transformer conditioned on pre-computed text tokens. It also introduces a tokenizer for learning video representation which compresses the video into discrete tokens. Using causal attention in time, it allows us to work with variable-length videos. MAGVIT [24] proposes an efficient video generation model through masked token modeling and multi-task learning. It first learns a 3D Vector Quantized (VQ) autoencoder to quantize videos into discrete tokens, and then learns a video transformer through multi-task masked token modeling. MAGVIT-v2 is a video tokenizer designed to generate concise and expressive tokens for both video and image generation using a universal approach. With this new tokenizer, the authors demonstrated that LLM outperforms diffusion models on standard image and video generation benchmarks, including ImageNet and Kinetics. VideoPoet [25] adopts a multi-modal Transformer architecture with a decoder-only structure. It uses the MAGVIT-v2 tokenizer to convert images and videos of arbitrary length into tokens, along with audio tokens and text embeddings, unifying all modalities into the token space. Subsequent operations are carried out in the

token Space, enabling the generation of coherent, high-action videos up to 10 seconds in length at once.

The families based on diffusion models are shown on the right side of Fig. 2. Video Diffusion Models (VDM) presents the first result on video generation using diffusion models by extending the image diffusion architecture. VDM employs a space-time factorized U-Net, jointly training on image and video data. It also introduces a conditional sampling technique for extending long and high-resolution videos spatially and temporally. Make-A-Video extends a T2I model to T2V with a spatiotemporally factorized diffusion model, removing the need for text-video pairs. It fine-tunes the T2I model for video generation, benefiting from effective model weight adaptation and improved temporal information fusion compared to VDM. Imagen Video [26] is a text-conditional video generation system that uses a cascade of video diffusion models. It incorporates fully convolutional temporal and spatial super-resolution models and a v-parameterization of diffusion models, enabling the generation of high-fidelity videos with a high degree of controllability and world knowledge. Runway Gen-1 [27] extends latent diffusion models to video generation by introducing temporal layers into a pre-trained image model and training jointly on images and videos. PYoCo [28] explores fine-tuning a pre-trained image diffusion model with video data, achieving substantially better performance with photorealism and temporal consistency. VideoCrafter [29], [30] introduces two diffusion models: the T2V model generates realistic and cinematic-quality videos, while the I2V model transforms an image into a video clip while preserving content constraints. EMU VIDEO [31] generates images conditioned on the text and then generates videos conditioned on the text and generated image, using adjusted noise schedules and multi-stage training for high-quality, high-resolution video generation without a deep cascade of models. Stable Video Diffusion [32] is a latent video diffusion model that emphasizes the importance of a well-curated pre-training dataset, providing a strong multi-view 3D-prior for fine-tuning multi-view diffusion models that generate multiple views of objects. Lumiere [33] is a T2V diffusion model with a Space-Time U-Net architecture that generates the entire temporal duration of a video at once, leveraging spatial and temporal down- and up-sampling and a pre-trained text-to-image diffusion model to generate full frame-rate, low-resolution videos on multiple space-time scales.

The center of Fig. 2 shows the fusion of the language model and the diffusion model [17], which is believed as the way leading T2V to the SOTA. Video Diffusion Transformer (VDT) [34] is the pioneer in the fusion of transformer and diffusion model, demonstrating its enormous potential in the field of video generation. VDT's strength lies in its outstanding ability to capture temporal dependencies, enabling it to generate temporally coherent video frames, including simulating the physical dynamics of three-dimensional objects over time. The proposed unified spatiotemporal masking mechanism allows VDT to handle various video generation tasks, achieving wide applicability. VDT's flexible handling of conditional information, such as simple token space concatenation, effectively unifies information of different lengths and modalities. Unlike U-Net, which is primarily designed for images, Transformer can better handle the time dimension by leveraging its powerful tokenization and attention mechanisms to capture long-term or irregular temporal dependencies. Only when the model learns (or memorizes) world knowledge, such as spatial-temporal relationships and physical laws, can it generate videos that match the real world. Therefore, the model's capacity becomes a key component of video diffusion. Transformer has proven to be highly scalable, making it more suitable than 3D U-Net for addressing the challenges of video generation. In December 2023, Stanford and Google introduced W.A.L.T [35], a transformer-based approach for latent video diffusion models (LVDMs), featuring two main design choices. Firstly, it employs a causal encoder to compress images and videos into a single latent space, facilitating cross-modality training and generation. Secondly, it utilizes a window attention architecture specifically designed for joint spatial and spatiotemporal generative modeling. This study represents the initial successful empirical validation of a transformer-based framework for concurrently training image and video latent diffusion models.

Sora's highlight is just the beginning of a new era in video generation, and it's foreseeable that this track will become very crowded. IT giants including Google, Microsoft, Meta, Baidu, startups like Runway, Pika, MidJourney, Stability.ai, as well as universities such as Stanford, Berkeley, Tsinghua, etc., are all powerful competitors.

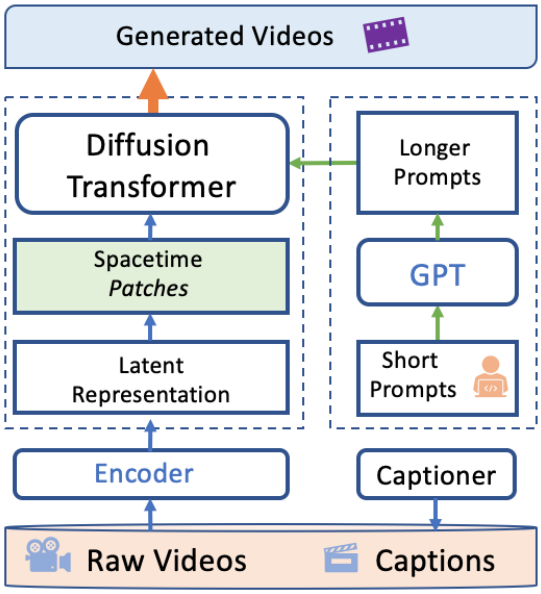

Fig. 3. Brief principle diagram of Sora.

Looking into Sora: A Parallel Intelligence Viewpoint

Upon its release, Sora sparked a huge wave of excitement, with its accompanying demos showcasing impressive results. Sora shows videos with high fidelity, rich details, significant object changes, and smooth transitions between multiple perspectives. While most video generation models can only produce videos lasting 3 to 5 seconds, Sora can create videos up to one minute in length while maintaining narrative

coherence, consistency, and common sense. Sora represents a milestone advancement in AI following ChatGPT.

What underpins Sora's powerful video generation capabilities? From Sora's technical report and the development history of video generation models, several key points can be summarized.

The first is the model architecture. Sora adopts the Diffusion Transformer (DiT), as shown in the left-upper corner Fig 3. Transformers have demonstrated powerful capabilities in large language models, with their attention mechanism effectively modeling long-range dependencies in spatiotemporal sequential data. Unlike earlier methods that perform windowed attention calculations or the Video Diffusion Transformer (VDT) that computes attention in the temporal and spatial dimensions separately, Sora merges the time and space dimensions and processes them through a single attention mechanism. Moreover, Transformers exhibit high computational efficiency and scalability, forming the basis for the Scaling Law with large models. The Diffusion Model, on the other hand, with its solid foundation in probability theory, offers high-resolution and good generation quality, as well as flexibility and controllability in video generation processes conditioned on text or images. DiT combines the advantages of both the Transformer and the Diffusion Model.

The second is data processing. As shown on the right side of Fig. 3, Sora leverages existing tools, such as the captioner used in DALL-E 3, to generate high-quality captions for raw videos, addressing the lack of video-text pairs. Additionally, through GPT, it expands users' short prompts to provide more precise conditions for video generation over long periods.

The third is feature representation. During training, Sora first compresses videos into a low-dimensional Latent Space (shown in the dashed rectangle on the left of Fig. 3) in both the spatial and temporal dimensions. Corresponding to the tokenization of text, Sora patchifies the low-dimensional representation in Latent Space into SpaceTime Patches, which are input into DiT for processing and ultimately generating new videos. From the perspective of parallel intelligence [36]–[44], the original videos come from the real system, while the Latent Space is the virtual system. Operations on the virtual system are more convenient to take advantage of the Transformer and the Diffusion Model.

Since OpenAI has not publicly disclosed the technical details of Sora, there may be other undisclosed technologies that have contributed to Sora's breakthrough in video generation capabilities. It should be noted that Sora's technical roadmap is far from mature. A large number of institutions are actively exploring and collaborating with each other. Microsoft, Google, Runway, Pika, Stanford, etc. have all iterated multiple versions and are still moving forward. The era of Imaginative Intelligence is just beginning.

Is Sora a World Model?

Although the released video clips from Sora have attracted a lot of attentions, OpenAI's claim that Sora is essentially a World Simulator or a World Model has sparked considerable controversy. Among them, LeCun's criticism is the most noteworthy.

A world model is a system that comprehends the real world and its dynamics. By using various types of data, it can be trained in an unsupervised manner to learn a spatial and temporal representation of the environment, in which we can simulate a wide range of situations and interactions encountered in the real world. To create these models, researchers face several open challenges, such as keeping consistent maps of the environment and the ability to navigate and interact within it. A world model must also capture not just the dynamics of the world but also the dynamics of its inhabitants, including machines and humans.

Thus, can Sora be called a world model? We analyze this from two perspectives.

Firstly, has Sora learned a world model? From the output results, most video clips are smooth and clear, without strange or jumpy scenes, and they align well with common sense. Sora can generate videos with dynamic camera movements. As the camera moves and rotates, characters and scene elements move consistently in a 3D environment. This implies that Sora already has the potential to understand and create in Spatial-Temporal space. Through these official demos, some have exclaimed that Sora has blurred the boundaries between reality and virtual for the first time in history. Therefore, we can say that Sora has learned some rules of real-world dynamics. However, upon closer observation of these videos, there are still some scenes that violate the laws of reality. For example, the process of a cup breaking, the incorrect direction of a treadmill, a puppy suddenly appearing and disappearing, ants having only four legs, etc. This indicates that Sora still has serious knowledge flaws in complex scenes, time scales, etc. There is still a significant gap compared to a sophisticated physics engine.

Secondly, does Sora represent the direction of world model development? From a technical perspective, Sora combines the advantages of large language models and diffusion models, representing the highest level of generative models. Scaling video generation models like Sora seems to be a promising approach to build a universal simulator for the physical world, which is a key step toward AGI. However, Yann LeCun has a different view. He believes that generative models need to learn the details of every pixel, making them too inefficient and doomed to fail. As an advocate for world models, he led Meta's team to propose Joint Embedding Predictive Architecture (JEPA) [45], believing that predictive learning in joint embedding space is more efficient and closer to the way humans learn. The latest release of V-JEPA also demonstrates the preliminary results of this approach.

In summary, Sora has gained a certain understanding of real-world dynamics. However, its functionality is still very limited, and it struggles with complex scenarios. Whether Sora ultimately succeeds or fails, it represents a meaningful attempt on the road to exploring World Models. Other diverse technical paths should also be encouraged.

Impacts

Sora and other video generation models have opened up new horizons for Imaginative Intelligence. PGC (Professional Generated Content) will widely adopt AI tools for production,

while UGC (User Generated Content) will gradually be replaced by AI tools. This commercialization of AI-generated video tools will accelerate, profoundly impacting various social domains. In fields like advertising, social media, and short videos, AI-generated videos are expected to lower the barrier to short video creation and improve efficiency. Sora also has the potential to change traditional film production processes by reducing reliance on physical shooting, scene construction, and special effects, thereby lowering film production costs. Additionally, in the field of autonomous driving [46], [47], Sora's video generation capabilities can provide training data, addressing issues such as data long-tail distribution and difficulty in obtaining corner cases [12].

On the other hand, Sora has also brought about social controversies. For example, Sora has raised concerns about the spread of false information. Its powerful image and video generation capabilities reach a level of realism that can deceive people, changing the traditional belief of "seeing is believing," making it harder to verify the authenticity of video evidence. The use of AI to forge videos for fraud and spread false information can challenge government regulation and lead to social unrest. Furthermore, Sora may lead to copyright disputes, as there could be potential infringement risks even in the materials used during the training process. Some also worry that generated videos could exacerbate religious and racial issues, intensifying conflicts between different religious groups, ethnicities, and social classes.

TAO to the Future of Imaginative Intelligence

Imaginative Intelligence. On the path to achieving imaginative intelligence, Sora represents a significant leap forward in AI's ability to visualize human imagination on a plausible basis. Imaginative intelligence, the highest level of the three layers of intelligence, goes beyond learning data, understanding texts, and reasoning. It deals with high-fidelity visual expressions and intuitive representations of imaginary worlds. After ChatGPT made advances in linguistic intelligence through superior text comprehension and logical reasoning, Sora excels at transforming potential creative thoughts into visualized scenes, giving AI the ability to understand and reproduce human imagination. This achievement not only provides individual creators with a quick way to visualize imaginary worlds, but also creates a conducive environment for collective creativity to collide and merge. It overcomes language barriers and makes it possible to merge ideas from different origins and cultures on a single canvas and ignite new creative inspirations. Sora has the potential to be a groundbreaking tool for humanity, allowing exploration of unknown territories and prediction of future trends in virtual environments. As technology continues to advance and its applications expand, the development of Sora and analog technologies signals the beginning of a new era in which human and machine intelligence reinforce each other and explore the boundaries of the imaginary world together.

Scenarios Engineering plays a crucial role in promoting the smooth and secure operation of artificial intelligence systems. It encompasses various processes aimed at optimizing the environment and conditions in which artificial intelligence operates, thereby maximizing its efficiency and safety [48]–[51]. With the emergence of advanced models like Sora, which specialize in converting text inputs into video outputs, not only new pathways for generating dynamic visual content are provided but also the capabilities of Scenarios Engineering are significantly enhanced [52]–[54]. This, in turn, contributes to the improvement of intelligent algorithms through enhanced calibration, validation, analysis, and other fundamental tasks.

Blockchain and Federated Intelligence. In its very essence, blockchain technology serves to underpin and uphold the "TRUE" characteristics, standing for trustable, reliable, usable, and effective/efficient [55]. Federated control is achieved based on blockchain technology, supporting federated security, federated consensus, federated incentives, and federated contracts [56]. Federated security comes from the security mechanism in the blockchain, playing a crucial role in the encryption, transmission, and verification of federated data [57]. Federated consensus ensures distributed consensus among all federated nodes on strategies, states, and updates. Federated incentives in federated blockchain are established for maintenance and management [58]. Therefore, designing fast, stable, and positive incentives can balance the interests between federated nodes, stimulate the activity of federated nodes, and improve the efficiency of the federated control system. Federated contracts [59] are based on smart contract algorithms that automatically and securely implement federated control. Federated contracts mainly function in access control, non-private federated data exchange, local and global data updates, and incident handling.

DeSci and DAO/TAO. The emergence of new ideas and technologies presents great opportunities for paradigm innovation. For example, the wave of decentralized science (DeSci) is changing the way scientific research is organized. As AI research enters rapid iteration, there are calls to establish new research mechanisms to overcome challenges such as the lack of transparency and trust in traditional scientific cooperation, and to achieve more efficient and effective scientific discoveries. DeSci aims to create a decentralized, transparent, and secure network for scientists to share data, information, and research findings. The decentralized nature of DeSci enables scientists to collaborate more fairly and democratically. DAO, as a means of implementing DeSci, provides a new organizational form for AI innovation and application [60], [61]. DAO represents a digitally-native entity that autonomously executes its operations and governance on a blockchain network via smart contracts, operating independently without reliance on any centralized authority or external intervention [62]–[64]. The unique attributes of decentralization, transparency, and autonomy inherent in DAOs provide an ideal ecosystemic foundation for developing imaginative intelligence. However, practical implementation has also shed light on certain inherent limitations associated with DAOs, such as power concentration, high decision-making barrier, and the instability of value system [65]. As such, TRUE autonomous organizations and operations (TAO) were proposed to address these issues, by highlighting their fundamental essence of being "TRUE" instead of emphasizing the decentralized attribute of DAOs [66]. Within the TAO framework, decision-making processes are hinged upon community consensus, and resource allocation

follows transparent and equitable rules, thereby encouraging multidisciplinary experts and developers to actively engage in complex and cutting-edge AI development. Supported by blockchain intelligence [67], TAO stimulates worldwide interest and sustained investment in intelligent technologies by devising innovative incentive mechanisms, reducing collaboration costs and enhancing flexibility and responsiveness of community management. As such, TAO provides an ideal ecosystem for nurturing, maturing, and scaling up the development of groundbreaking technologies of imaginative intelligence.

When will Sora or Sora-like AI Technology show us the real road or TAO to Imagitative Intelligence that could be practically used for constructing a sustainable and smart society with intelligent industries for better humanity? We are still expecting, but more enthusiastically now.

ACKNOWLEDGMENT

This work was partially supported by the National Natural Science Foundation of China (62271485, 61903363, U1811463, 62103411, 62203250) and the Science and Technology Development Fund of Macau SAR (0093/2023/RIA2, 0050/2020/A1).

REFERENCES

[1] F.-Y. Wang *et al.*, "What Does ChatGPT Say: The DAO from Algorithmic Intelligence to Linguistic Intelligence," *IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 3, pp. 575–579, 2023.

[2] F.-Y. Wang *et al.*, "What Does MovieGIIT Show: Intelligent Science and Technology after AlphaGo and ChatGPT," *Journal of Intelligence Science and Technology*, vol. 3, no. 1, pp. 1–6, 2023.

[3] F.-Y. Wang, "Foundation worlds for parallel intelligence: From foundation/infrastructure models to foundation/infrastructure intelligent," *Alfred North Whitehead Laureate Lecture*, TWAI, Beijing, Nov. 2021.

[4] F.-Y. Wang, "Parallel directors for decision theaters: Decision intelligence for new sciences," *Journal of Intelligence Science and Technology*, vol. 3, no. 1, pp. 7–12, 2023.

[5] Q. Ni *et al.*, "Parallel theaters: human-machine collaborative creation and intelligent management for theatrical art," *Chinese Journal of Intelligent Science and Technology*, vol. 5, no. 4, pp. 436–445, 2023.

[6] Y. Tian *et al.*, "VistaGPT: Generative parallel transformers for vehicles with intelligent systems for transport automation," *IEEE Transactions on Intelligent Vehicles*, 2023.

[7] F.-Y. Wang *et al.*, "Where does AlphaGo go: From Church-Turing Thesis to AlphaGo Thesis and beyond," *IEEE/CAA Journal of Automatica Sinica*, vol. 3, no. 2, pp. 113–120, 2016.

[8] F.-Y. Wang, "Intelligent industries and industrial revolution: The imaginative DAO to continuous ownership and new economics of parallel supply-demand ecology," *PEW 2020*, 2020.

[9] F.-Y. Wang, "DeSci for DeDAO in CPSS: Parallel intelligence and intelligent industries for smart societies," *DHW on DeSci*, 2022.

[10] F.-Y. Wang, "Parallel management: The DAO to smart ecological technology for complexity management intelligence," *Acta Automatica Sinica*, vol. 48, no. 11, pp. 2655–2669, 2022.

[11] W. Ding et. al., "Desci based on Web3 and DAO: A comprehensive overview and reference model," *IEEE Transactions on Computational Social Systems*, vol. 9, no. 5, pp. 1563–1573, 2022.

[12] X. Li *et al.*, "Sora for senarios engineering of intelligent vehicles: V&V, C&C, and beyonds," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 2, pp. 3025–3029, 2024.

[13] R. Qin *et al.*, "Sora for computational social systems: From counterfactual experiments to artificiofactua experiments," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 2, pp. 1529–1548, 2024.

[14] L. Fan *et al.*, "Sora for foundation robots with parallel intelligence: Three world models, three robotic systems," *Frontiers of Information Technology Electronic Engineering*, March, 2024.

[15] H. Yu *et al.*, "Sora-based parallel vision for smart sensing of intelligent vehicles: from foundation models to foundation intelligence," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 2, pp. 3030–3033, March, 2024.

[16] J. Li *et al.*, "Digital CEOs in digital enterprises: Automating, augmenting, and parallel in Metaverse/CPSS/TAOs," *IEEE/CAA Journal of Automatica Sinica*, vol. 11, no. 4, pp. 820–823, 2024.

[17] W. Peebles *et al.*, "Scalable Diffusion Models with Transformers," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Paris, France: IEEE, Oct. 2023, pp. 4172–4182.

[18] W. Yan *et al.*, "Videogpt: Video generation using vq-vae and transformers," *arXiv preprint arXiv:2104.10157*, 2021.

[19] C. Wu *et al.*, "NÜWA: Visual Synthesis Pre-training for Neural visUal World creAtion," in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, vol. 13676, pp. 720–736.

[20] W. Hong *et al.*, "Cogvideo: Large-scale pretraining for text-to-video generation via transformers," *arXiv preprint arXiv:2205.15868*, 2022.

[21] A. Gupta *et al.*, "Maskvit: Masked visual pre-training for video prediction," 2022.

[22] S. Ge et. al., "Long video generation with time-agnostic vqgan and time-sensitive transformer," *arXiv preprint arXiv:2204.03638*, 2022.

[23] R. Villegas *et al.*, "Phenaki: Variable length video generation from open domain textual description," 2022.

[24] L. Yu *et al.*, "MAGVIT: Masked generative video transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[25] D. Kondratyuk *et al.*, "Videopoet: A large language model for zero-shot video generation," 2023.

[26] J. Ho *et al.*, "Imagen Video: High Definition Video Generation with Diffusion Models," Oct. 2022.

[27] E. Patricket *et al.*, "Structure and content-guided video synthesis with diffusion models," 2023.

[28] S. Ge *et al.*, "Preserve Your Own Correlation: A Noise Prior for Video Diffusion Models," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Paris, France: IEEE, Oct. 2023, pp. 22 873–22 884.

[29] H. Chen *et al.*, "Videocrafter1: Open diffusion models for high-quality video generation," 2023.

[30] H. Chen *et al.*, "Videocrafter2: Overcoming data limitations for high-quality video diffusion models," 2024.

[31] R. Girdhar *et al.*, "Emu video: Factorizing text-to-video generation by explicit image conditioning," 2023.

[32] A. Blattmann *et al.*, "Stable video diffusion: Scaling latent video diffusion models to large datasets," 2023.

[33] O. Bar-Tal *et al.*, "Lumiere: A space-time diffusion model for video generation," 2024.

[34] H. Lu *et al.*, "Vdt: General-purpose video diffusion transformers via mask modeling," 2023.

[35] A. Gupta *et al.*, "Photorealistic video generation with diffusion models," 2023.

[36] L. Li, Y. Lin, N. Zheng *et al.*, "Parallel learning: A perspective and a framework," *IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 3, pp. 389–395, 2017.

[37] F.-Y. Wang *et al.*, "Steps toward parallel intelligence," *IEEE/CAA Journal of Automatica Sinica*, vol. 3, no. 4, pp. 345–348, 2016.

[38] P. Ye *et al.*, "Parallel cognition: Hybrid intelligence for human-machine interaction and management," *Frontiers of Information Technology & Electronic Engineering*, vol. 23, no. 12, pp. 1765–1779, 2022.

[39] J. Lu *et al.*, "Parallel factories for smart industrial operations: From big AI models to field foundational models and scenarios engineering," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 12, pp. 2079–2086, 2022.

[40] Y. Shen *et al.*, "Parallel sensing in metaverses: Virtual-real interactive smart systems for "6S" sensing," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 12, pp. 2047–2054, 2022.

[41] J. Yang *et al.*, "Parallel manufacturing for industrial metaverses: A new paradigm in smart manufacturing," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 12, pp. 2063–2070, 2022.

[42] F.-Y. Wang, "Parallel intelligence in metaverses: Welcome to Hanoi!" *IEEE Intelligent Systems*, vol. 37, no. 1, pp. 16–20, 2022.

[43] Q. Miao *et al.*, "Parallel learning: Overview and perspective for computational learning across Syn2Real and Sim2Real," *IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 3, pp. 603–631, 2023.

[44] Z. Song *et al.*, "Parallel learning for legal intelligence: A hanoi approach based on unified prompting," *IEEE Transactions on Computational Social Systems*, pp. 1–11, 2023.

[45] A. Bardes *et al.*, "Revisiting feature prediction for learning visual representations from video," *arXiv preprint*, 2024.
[46] S. Teng *et al.*, "Motion planning for autonomous driving: The state of the art and future perspectives," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 6, pp. 3692–3711, 2023.
[47] C. Zhao *et al.*, "Decast in transverse for parallel intelligent transportation systems and smart cities: Three decades and beyond," *IEEE Intelligent Transportation Systems Magazine*, vol. 14, no. 6, pp. 6–17, 2022.
[48] Li, Xuan *et al.*, "From features engineering to scenarios engineering for trustworthy ai: I&i, c&c, and v&v," *IEEE Intelligent Systems*, vol. 37, no. 4, pp. 18–26, 2022.
[49] X. Li *et al.*, "A novel framework to generate synthetic video for foreground detection in highway surveillance scenarios," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 6, pp. 5958–5970, 2023.
[50] X. Li *et al.*, "A novel scenarios engineering methodology for foundation models in metaverse," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 53, no. 4, pp. 2148–2159, 2023.
[51] X. Wang *et al.*, "Smart decentralized autonomous organizations and operations for smart societies: Human–autonomous organizations for industry 5.0 and society 5.0," *IEEE Intelligent Systems*, vol. 38, no. 06, pp. 70–74, nov 2023.
[52] X. Li *et al.*, "Development and testing of advanced driver assistance systems through scenario-based systems engineering," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 8, pp. 3968–3973, 2023.
[53] X. Li *et al.*, "Advanced scenario generation for calibration and verification of autonomous vehicles," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 5, pp. 3211–3216, 2023.
[54] Li, Xuan *et al.*, "A novel scenarios engineering methodology for foundation models in metaverse," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 53, no. 4, pp. 2148–2159, 2023.
[55] J. Li *et al.*, "The Future of Management: DAO to Smart Organizations and Intelligent Operations," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 53, no. 6, pp. 3389–3399, 2023.
[56] F.-Y. Wang *et al.*, "Federated Control: Toward Information Security and Rights Protection," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 4, pp. 793–798, Aug. 2021.
[57] F.-Y. Wang *et al.*, "Federated Ecology: Steps Toward Confederated Intelligence," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 2, pp. 271–278, Apr. 2021, conference Name: IEEE Transactions on Computational Social Systems.
[58] F.-Y. Wang *et al.*, "Federated Management: Toward Federated Services and Federated Security in Federated Ecology," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 6, pp. 1283–1290, Dec. 2021.
[59] S. Wang *et al.*, "Blockchain-enabled smart contracts: Architecture, applications, and future trends," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 11, pp. 2266–2277, 2019.
[60] Q. Miao *et al.*, "Dao to HANOI via DeSci: AI Paradigm Shifts from AlphaGo to ChatGPT," *IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 4, pp. 877–897, 2023.
[61] C. Zhao *et al.*, "Decentralized autonomous operations and organizations in transverse: Federated intelligence for smart mobility," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 53, no. 4, pp. 2062–2072, 2023.
[62] R. Qin *et al.*, "Web3-based decentralized autonomous organizations and operations: Architectures, models, and mechanisms," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 53, no. 4, pp. 2073–2082, 2023.
[63] J. Li, R. Qin, W. Ding, G. Wang, T. Wang, and F.-Y. Wang, "A new framework for web3-powered decentralized autonomous organizations and operations," *Acta Automatica Sinica*, vol. 49, no. 5, pp. 985–998, 2023.
[64] W. Ding *et al.*, "A novel approach for predictable governance of decentralized autonomous organizations based on parallel intelligence," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 53, no. 5, pp. 3092–3103, 2023.
[65] J. Li *et al.*, "The TAO of Blockchain Intelligence for Intelligent Web 3.0," *IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 12, pp. 2183–2186, 2023.
[66] J. Li, X. Liang, R. Qin, and F.-Y. Wang, "From DAO to TAO: Finding The Essence of Decentralization," in *The IEEE International Conference on Systems, Man, and Cybernetics*, Hawaii, Oct. 1-4, 2023.
[67] J. Li, R. Qin, S. Guan, J. Hou, and F.-Y. Wang, "Blockchain intelligence: Intelligent blockchains for web 3.0 and beyond," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pp. 1–10, 2024.

ABOUT THE AUTHOR

Fei-Yue Wang (Fellow, IEEE) Bio of Fei-Yue Wang can be found at https://ieeexplore.ieee.org/author/37277656000.

Qinghai Miao (Senior Member, IEEE) Bio of Qinghai Miao can be found at https://ieeexplore.ieee.org/author/37685880900.

Lingxi Li (Senior Member, IEEE) Bio of Lingxi Li can be found at https://ieeexplore.ieee.org/author/37293042500.

Qinghua Ni Bio of Qinghua Ni can be found at https://ieeexplore.ieee.org/author/37090024499.

Xuan Li Bio of Xuan Li can be found at https://ieeexplore.ieee.org/author/37085417532.

Juanjuan Li (Member, IEEE) Bio of Juanjuan Li can be found at https://ieeexplore.ieee.org/author/38468375000.

Lili Fan Bio of Lili Fan can be found at https://ieeexplore.ieee.org/author/37089635745.

Yonglin Tian Bio of Yonglin Tian can be found https://ieeexplore.ieee.org/author/37086460166.

Qing-Long Han (Fellow, IEEE) Bio of Qing-Long Han can be found at https://ieeexplore.ieee.org/author/37274962400.